Clustering and Staircases

K. Pelckmans, J.A.K. Suykens, B. De Moor
K.U.Leuven - ESAT - SCD, Leuven - Belgium,
MINE - IPSI - Fraunhofer, Darmstadt, Germany,
kristiaan.pelckmans@esat.kuleuven.ac.be

Abstract

Clustering[1] denominates a range of different tasks including vector quantization, graph-cut problems, bump-hunting and optimal compression. This presentation motivates the viewpoint that the class of staircases is implicitly the object of study underlying those various tasks. This assertion provides a natural way to formulate and study the theoretical counterpart to many empirical clustering algorithms.

Clustering Shrinkage

Empirical clustering shrinkage was studied in (Pelckmans *et al.*, 2005)[2], following from a costfunction for a set of unlabeled datapoints $\mathcal{D} = \{x_i\}_{i=1}^N \subset \mathbb{R}^D$ and its corresponding representatives (or centroids) presented in functional form as $\mathcal{M} = \{m(x_i)\}_{i=1}^N \subset \mathbb{R}^D$:

$$\hat{m} = \underset{m:\mathbb{R}^D \to \mathbb{R}^D}{\arg\min} \mathcal{J}_\gamma^{p,q}(m) = \frac{1}{p}\sum_{i=1}^N \|m(x_i) - x_i\|_p + \gamma \sum_{i<j} \|m(x_i) - m(x_j)\|_q , \tag{1}$$

where attention is restricted to the projection of the function $\hat{m}$ to the datapoints $\mathcal{M}$. We denote the two terms of the rhs. as the *reconstruction term* and *the clustering regularization term* respectively. Note that sparseness in the difference between two centroids $\|m(x_i) - m(x_j)\| = 0$ indicate that the corresponding datapoints x_i and x_j are assigned to a common cluster with centroid $m(x_i) = m(x_j)$. The mentioned paper studied the convex counterpart using $p = 2$ and $q = 1$ (cfr. the LASSO estimator) which can be solved as a QP problem using standard software tools.

This presentation will focus on the consequences of the clear optimization point of view from a theoretical perspective. First of all, we consider the case where we count the number of nonzero differences (informally denoted as $q = 0$). It was argued that the resulting costfunction is minimized by a k-means algorithm using an alternating global optimization algorithm. A second improvement to (1) shifts the focus to local differences instead of the global term $\sum_{i<j} \|m(x_i) - m(x_j)\|$:

$$\hat{m}_\epsilon = \underset{m:\mathbb{R}^D \to \mathbb{R}^D}{\arg\min} \mathcal{J}_\gamma^{\epsilon,p}(m) = \frac{1}{p}\sum_{i=1}^N \|m(x_i) - x_i\|_p + \frac{\gamma}{|B(\epsilon)|}\sum_{i=1}^N \sum_{\|x_i - x_j\| \leq \epsilon} I\left(\|m(x_i) - m(x_j)\| > 0\right), \tag{2}$$

where $|B(\epsilon)|$ measures the volume of the balls $B(\epsilon; x) = \{y \in \mathbb{R}^D : \|x - y\| \leq \epsilon\}$ with radius ϵ. As such, the second term measures the density of different assigned datapoints in a local neighborhood employing a similar mechanism as in the histogram density estimator. Note that the case where $\epsilon \to +\infty$ correspond with (1) where $q = 0$. If $\epsilon \to 0$ when $N \to +\infty$, the algorithm implementing (2) can be expected to converge to the following minimizer.

Definition 1 (Theoretical Shrinkage Clustering) *Let $m : \mathbb{R} \to \mathbb{R}$ be such that $\lim_{\|\delta\| \to 0} \frac{m(x-\delta) - m(x+\delta)}{|B(\|\delta\|)|}$ exists almost everywhere. Let the cdf $P(x)$ underlying the dataset be known and assume its pdf $p(x)$ exists everywhere and is nonzero on a connected compact interval $C \in \mathbb{R}$ with nonzero measure $|C| > 0$. We*

[1]The authors like to acknowledge constructive discussions with U. von Luxburg, J. Shawe-Taylor, M.Pontil, O. Chapelle, A.Zien and others.

[2]K. Pelckmans J.A.K. Suykens and B. De Moor, Convex clustering shrinkage. In *"Statistics and Optimization of Clustering Workshop", PASCAL*, 2005.

will study the following theoretical counterpart to (2)

$$\hat{m} = \underset{m:\mathbb{R}\to\mathbb{R}}{\arg\min}\, \mathcal{J}_\gamma^{p,0}(m) = \int_C \left\|m(x) - x\right\|_p dP(x) + \gamma \int_C \left\|m'(x)\right\|_0 dP(x), \tag{3}$$

where we define the latter term -denoted further as the zero-norm variation- formally as follows

$$\left\|m'(x)\right\|_0 \triangleq \lim_{\epsilon\to 0}\left(\frac{I\left(m(B(x;\epsilon)) \neq const\right)}{|B(x,\epsilon)|}\right), \tag{4}$$

with the characteristic function $I\big(m(B(x;\epsilon)) \neq const\big)$ equals one if $\exists y \in B(x;\epsilon)$ such that $\|m(x) - m(y)\| > 0$ ($B(x,\epsilon)$ contains parts of different clusters), and equal to zero otherwise.

Intuitively, the zero-norm variation expresses the probability that any point $x \in C$ cannot be assigned to the same cluster as its immediate neighbors, representing the required degenerate solutions (clusters). This construction triggers the following representer result. We restrict for this presentation the attention to the univariate case $D = 1$ for notational convenience.

Theorem 1 (Univariate Staircase Representation) *When $P(x)$ is a fixed, smooth and differentiable distribution function with pdf $p : \mathbb{R} \to \mathbb{R}^+$ which is nonzero on a compact interval $C \subset \mathbb{R}$, the minimizer to (3) takes the form of a staircase function uniquely defined on C with a finite number of positive steps (say $K < +\infty$) of size $a = (a_1, \ldots, a_K)^T \in \mathbb{R}^K$ at the points $\mathcal{D}_{(K)} = \{x_{(k)}\}_{k=1}^K \subset C$*

$$\hat{m}\left(x; a, \mathcal{D}_{(K)}\right) = \sum_{k=1}^{K} a_k I\left(x > x_{(k)}\right) \quad s.t. \quad a_k \geq 0, x_{(k)} \in C \quad \forall k \tag{5}$$

Moreover, the optimization problem (3) is equivalent to the problem

$$\min_{a,\mathcal{D}_{(K)}} \mathcal{J}_K^p\left(a, \mathcal{D}_{(K)}\right) = \int_C \left\|\sum_{k=1}^{K} a_k I\left(x > x_{(k)}\right) - x\right\|_p p(x)dx + \sum_{k=1}^{K} p(x_{(k)}), \tag{6}$$

where $K \in \mathbb{N}$ relates to $\gamma \in \mathbb{R}^+$ in a way depending on $\mathcal{D}$.

The proof is based on the fact that the term (6) grows unboundedly if m has an input region in C - say the region $[a, b] \subset C$ with $a < b$ - where the function is nonconstant such that $m'(x) > 0$ for $a \leq x \leq b$. From this it follows that the following inequality holds

$$\int_C \left\|m'(x)\right\|_0 dP(x) \geq \lim_{0<\delta\to 0} \int_a^b \left(\frac{I\big(|m(x-\delta) - m(x+\delta)| > 0\big)}{2\delta}\right) dP(x) \geq \lim_{0<\delta\to 0} \int_a^b \frac{dP(x)}{2\delta} \to +\infty \tag{7}$$

and the zero-norm variation becomes unbounded whenever the function m contains not only variations on sets with zero measure (steps). Monotonicity of a_k follows directly from the reconstruction term.

Interpretations

Equation (6) then underlies various techniques collected under the denominator of clustering. While vector quantization algorithms as k-means **(I)** emphasize the reconstruction term, In density based algorithms -also referred to as *bump-hunting* - and min-cut algorithms **(II)**, the regularization term is stressed while the reconstruction term keeps the cut normalized. Moreover, it it is argued that by considering the solution-path $\mathcal{S} = \left\{\hat{M} \mid \exists\gamma \quad \text{s.t.} \quad \hat{M} = \arg\min_M \mathcal{J}_\gamma^{p,q}(M)\right\}$, one obtains the result of an hierarchical clustering algorithm **(III)**. Furthermore one may also view (6) as approaching the task of optimal coding **(IV)**, in the sense of "finding a short code for X that preserves the maximum information about X itself." Note that by replacing the reconstruction term by the KL-distance between X and $m(X)$, a more information theoretic oriented context can be adopted. Finally, we want to hint to the problem of finding the optimal bin placement of a histogram **(V)** for optimally reconstructing the density underlying a finite dataset. This link can play an important role in the task of histogram based density estimation based on multivariate data (e.g. $D = 3, 4$).

An important consequence of Theorem 1 is that analysts now have to study the class of staircase functions (as in e.g. classification), its projection on the given dataset $\mathcal{D}$ (cfr. assignment problem), and the evaluation of the staircase in new points (cfr. extension operator). This discriminates this track from the research on (local) convergence of proposed algorithms and gives a clearcut interpretation of the notion of stability (regularization) in clustering algorithms. [3]

[3]- (KP): BOF PDM/05/161, FWO grant V4.090.05N; - (SCD:) GOA AMBioRICS, CoE EF/05/006, (FWO): G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.0553.06, (ICCoS, ANMMM, MLDM); (IWT): GBOU (McKnow), Eureka-Flite2 IUAP P5/22,PODO-II,FP5-Quprodis; ERNSI; - (JS) is an associate professor and (BDM) is a full professor at K.U.Leuven Belgium, respectively.